A Framework for Infection Control Surveillance Using Association Rules

Lili Ma[1, 2], Fu-Chiang Tsui PhD[1], William R. Hogan MD[1], Michael M. Wagner MD, PhD[1], Haobo Ma MD[1]

[1]RODS Laboratory, Center of Biomedical Informatics and [2]Intelligent Systems Program,
University of Pittsburgh, Pittsburgh, PA

ABSTRACT

Surveillance of antibiotic resistance and nosocomial infections is one of the most important functions of a hospital infection control program. We employed the association rule method for automatically identifying new, unexpected, and potentially interesting patterns in hospital infection control. We hypothesized that mining for low-support, low-confidence rules would detect unexpected outbreaks caused by a small number of cases. To build a framework, we preprocessed the data and added new templates to eliminate uninteresting patterns.

We applied our method to the culture data collected over 3 months from 10 hospitals in the UPMC Health System. We found that the new process and system are efficient and effective in identifying new, unexpected, and potentially interesting patterns in surveillance data. The clinical relevance and utility of this process await the results of prospective studies.

INTRODUCTION

Nosocomial infections — also known as hospital-acquired infections — are infections that patients acquire during the course of their hospital stay. In developed countries, about 5 to 10% of patients acquire nosocomial infection, while in the developing countries the rate exceeds 25%. Such nosocomial infections result in unexpected morbidity, mortality, and additional costs to hospitals [1]. Each year in the United States nosocomial infections affect 2 million patients, cost more that $4.5 billion, and account for half of all major hospital complications [2]. Without early detection and proper control of infected patients, other patients or healthcare staff can be infected. Thus early recognition of outbreaks and emerging resistance requires proactive surveillance at the hospital and sub-hospital levels.

Conventionally, nosocomial infection surveillance has relied on ward rounds, reviews of medical charts and paper-based reports of microbiologic results. Analyses are typically conducted by assembling hand written data or manually entering information into a computer database. These conventional infection control methods are time consuming, labor intensive and relatively inefficient for quantitative analyses and coping with the increasing complexity of antibiotic resistance. The potential value of computer-based surveillance is widely recognized and recent reports have described effective computer applications for infection control [3].

Extensive analysis of hospital data, however, requires considerable time and resources, both of which few hospital epidemiologists have in reserve. Consequently, these data are underutilized and the patterns they contain go undiscovered [4]. Brossette presented a new method, association rule induction, to detect temporal patterns among infection-control surveillance data in reference [4].

Association rule induction is a powerful data mining method for finding temporal trends in large datasets. The goal of data mining is to automate the process of finding interesting temporal patterns. The output of a data-mining method should be a “summary” of the data sets. Such goal is difficult to achieve due to the vagueness of the term “interesting”. The solution is to define various types of trends (patterns) and to look for only those defined trends in the data sets. One such type of trend is the association rule.

A number of statistical strategies have been developed for automatically detecting temporal patterns in surveillance data. Historically, computer-assisted infection control surveillance research has focused on identifying high-risk patients including those on suboptimal antibiotic regimens, the use of expert systems to identify possible cases of nosocomial infection, and the detection of temporal trends in the occurrence of predefined events [4].

From our discussions with infection control practitioners, we discovered that one requirement for data mining is to discover unusual occurrences, rather than frequent, regularly-occurring patterns. Therefore, we designed an association rule method by extending Brosette’s approach [4] to assist traditional infection-control surveillance by automatically detecting temporal trends – for example, trends in antibiotic resistance – that would not have been detected by either traditional or existing computer-assisted surveillance systems.

The main problem of association rule induction is that the number of possible rules to search over is intractable. However we do not want just any association rule produced by the method; we want “good” rules that are “expressive” and “reliable” as defined by standard measures of the “goodness” or “reliability” of association rules (described in detail in *Methods*). Standard criteria are often not sufficient to restrict the set of rules to the interesting ones, especially when the thresholds for both measures are low. Efficient algorithms are needed to restrict the search space and check only a subset of all rules for temporal trends in the data, but if possible, without missing important rules. Therefore, we considered and evaluated some additional rule evaluation measures, including data preprocessing and the use of templates.

METHODS

Association Rules

An association rule expresses an association between (sets of) items, which may be products of a supermarket or a mail-order company, special equipment options of a car, optional services offered by telecommunication companies etc. For example, an association rule (variables' definitions can be found in figure 3) *hospital_unit=12S, drugcode=OXA, organismspecies=STAPHYLOCOCCUS, hospital=1 -> resultcode=R* states that if we pick a patient record at random and find out that it contained {*hospital_unit=12S, drugcode=OXA, organismspecies=STAPHYLOCOCCUS, hospital=1*}, we can be confident that it also contained {*resultcode=R*}. An example of such a rule over basket data might be that 98% of customers that purchase tires and auto accessories also get automotive services done; finding all such rules is valuable for cross-marketing and attached mailing applications.

Definitions

The following is a mathematical statement of the association rule method: let $\mathcal{I} = \{i_1, i_2 \ldots i_3\}$ be a set of items, *e.g.,* drug codes, organisms isolated, wards, *etc*. Let $\mathcal{D}$ be a set of transactions, *e.g.,* culture reports, where each transaction T is a set of items such that $T \subseteq \mathcal{I}$. We say that a transaction T *contains* X, where $X \subseteq T$. An *association rule* is a rule that states X associates with Y, denoted by $X \Rightarrow Y$, where $X, Y \subseteq \mathcal{I}$, and $X \cap Y = \emptyset$ *(empty set)*. We define *support* of rule $X \Rightarrow Y$ as the proportion of transactions in $\mathcal{D}$ with both X and Y, *i.e.*, $X \cup Y$. We define *confidence* of a rule $X \Rightarrow Y$ as *support*($X \cup Y$)/*support*(X). Support is the percentage of transactions that the rule can be applied to (or, alternatively, the percentage of transactions, in which it is correct). Confidence is the number of cases in which the rule is correct relative to the number of cases in which it is applicable (and thus is equivalent to an estimate of the conditional probability of the consequent of the rule given its antecedent). A valid rule must at least satisfy the two criteria: the support and confidence must be greater than the user-specified thresholds--minimum support and minimum confidence respectively. Furthermore, itemsets with minimum support are called *large itemsets*, and all others are called *small itemsets*. The left hand side of a rule is called the cause part, and the right hand side of the rule is called the result part.

The Apriori Algorithm

In this study, we chose the *apriori algorithm* [5] for association rule induction. This algorithm works in two steps. The first step determines the large itemsets that have at least the given minimum support (*i.e.,* occur at least in a given percentage of all transactions). In the second step association rules are generated from the large itemsets found in the first step. Usually the first step is more important, because it accounts for the greater part of the processing time.

In order to make it efficient, the apriori algorithm exploits the observation that the superset of a small itemset (*i.e.*, an itemset without minimum support) can not be a large itemset (with enough support) [6].

Figure 1 is a frame of the Apriori algorithm.

```
$L_1 = \{$Large 1-itemsets$\}$;
for ( $k = 2$; $L_{k-1} \neq \emptyset$; $k$++) do begin
    $C_k$ = apriori-gen($L_{k-1}$);        // new candidates
    forall transactions $t \in \mathcal{D}$ do begin
        $C_t$ = subset($C_k$, $t$);  // candidates contained in $t$
        forall candidate $c \in C_t$ do
            $c$.count++;
    end
    $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
end
Answer = $\cup_k L_k$;
```

Figure 1: Algorithm Apriori

The *apriori-gen* function takes as argument L_{k-1}, the set of all large (k-1)-itemsets. It returns a superset of the set of all large k-itemsets. The function works as follows. First, in the *join* step, join L_{k-1} with L_{k-1}. Next, in the *prune* step, we delete all itemsets $c \in C_k$ such that some (k - 1)-subset of c is not in L_{k-1} [5].

Event Capture

An *event* describes an interesting change in the confidence of an association rule over time.

We partitioned patients' records into time slices, according to the date when the information was recorded. Each time slice covers one month of records since monthly time slice often has enough samples to do association rule induction.

Each association rule generated is compared to a set of user-defined *rule templates* that describe "flavors" of interesting and uninteresting rules. Since rule templates contain domain knowledge, domain experts must handcraft them. In general, an expert usually has an idea of what types of rules are interesting, or may know of some types that are never interesting [4]. There are two types of rule templates: *include templates* and *exclude templates*. An association rule passes a set of templates if it satisfies at least one include template in the set and does not satisfy any exclude template in the set [4]. Association rule templates can be found in Brossette [4].

We took advantage of the templates developed by Brossette et al [4]. In addition to those templates, we also added the templates as shown in Figure 2. Based on these templates, every association rule must include *organismspecies* and *drugcode* attributes in their cause parts (left hand side of⇒), other wise the rule will be filtered out, since it is less interesting or not interesting at all to hospital infection control experts. For the result part (right hand side of⇒) of a rule, the *resultcode* must exist.

Type	*Left*		*Right*	*Explanation*
Exclude	*(resultcode)*	⇒	*(Anything)*	*Want test result information on the right only.*
Exclude	*(Anything)*	⇒	*(Anything except resultcode)*	*Want every rule must include resultcode as result*
Include	*(oganismcode* AND *drugcode)*	⇒	*(Anything)*	*Want every rule must include oganismcode and drugcode in left.*

Figure 2: Additional association rule templates

The *history* is a database that holds association rules and their information, like support, confidence, et al, for different data partitions. Only association rules that pass the rule templates are stored in the history list.

For each pair of time slices, a *chi-square-base test* is performed to determine whether there is a significant difference in the confidence of the rule between time slices. Association rules passing this test will be output as events.

Chi-square-base Test

The most common method of measuring strength of association is the calculation of χ^2 on one degree of freedom from the comparison of two binomials. The calculation is relatively simple and with count data (numbers of individuals exposed or not, numbers of individuals with the outcome or not) is identical for all fourfold tables, and it does not matter whether they originate from cohort studies, case-referent studies, or prevalence studies.

The conventional interpretation of these probabilities is that a *P* value of <0.05 indicates that the observed difference is unlikely to have occurred by chance alone and, thus, somehow must represent a real difference; another way of stating this is that we are 95% certain that this observed difference could not have arisen by chance alone.

Dataset—Microbiology reports

The Health System Resident Component (HSRC) located in the UPMC Health System was developed by the RODS laboratory starting in 1999. The HSRC receives microbiological HL7 messages in real time from 10 hospitals including 131 hospital units [7]. The study period was from May 1st, 2000 to July 31st, 2000. Within the study period, there were 941 microbiology transactions total. Patients' medical record number (MRN) was encrypted to protect patient privacy. The data set also included coded elements—hospital, ward, attending doctor, etc.

Data Preprocessing

Each patient has three duplicate reports: preliminary, intermediate and final report. We used three data elements as a unique key — patient's name, accession number and sampling date — to remove duplicates. We kept only the most recent records. We store the culture data in three tables: method, organism and patient. Attributes in these tables are: ecrypted_pid, gender, age, hospital, hospital_unit, attending_doctor, lab_id, isolateid, accession_number, sampling_time, date_spec_reported, drugname, dat_spec_received, organismcode, drugcode, resultcode, rultname, organismname, methodcode, methodname.

According to domain experts' suggestion, we categorized patients' ages into four categories: 0–10 years, 11–20 years, 21–60 years, 61 years and up. We grouped organisms at the level of species. We designed these groupings to reduce the amount of output without losing interesting information. We also eliminated several uninteresting attributes such as isolateid, drugname (duplicate information that is recorded in drugcode), etc. from the data set. We combined the three original tables into one transaction table.

The reason to choose May, June and July as study period is as follows: First a three month period will contain enough records to do association rule induction; second; second these three months belong to the same season, so we eliminate the influence of seasonal variations. We aggregated data by month since a monthly time slice often has enough samples to do association rule induction.

Figure 3 is a summary of the experiment dataset after data preprocessing.

Attributes in Culture Transaction Table	***Variable descriptions***	***Number of different values***
organismspecies	*Species of the organism*	*30*
drugcode	*Drugs used in organism's susceptibility test*	*48*
resultcode	*Results of organism's susceptibility test: Sensitive (S), Resistant (R), or Intermediate (I).*	*3*
gender	*Male or female*	*2*
agegroup	*1: age 0 to 10* *2:age 11 to 20* *3: age 21 to 60* *4: age 61 and above*	*4*
hospital	*Hospital ID*	*10*
hospital_unit	*Hospital ward*	*131*
attending_doctor	*Attending doctor*	*377*
month	*The month when the organism sensitive tests were performed.*	*3*
Date	*Specimen collected date*	*31*

Figure 3: Summary of Experiment Dataset after Data Preprocessing

The Infection Control Surveillance System

Traditional association rule data mining applications focus on discovering high-support, high-confidence association rules, for these rules can be used for classification. In reference [4], Brossette pointed out that while high-support, high-confidence rules will be useful in the surveillance paradigm, high-support, low-confidence rules would often be more useful. If B occurs every time A occurs, and A occurs frequently, then we maintain that the rule $A \Rightarrow B$ will probably be known or trivial and therefore uninteresting. However, if B occurs infrequently with A and A occurs relatively frequently, then $A \Rightarrow B$ is a low-confidence association rule, and changes in the confidence of $A \Rightarrow B$ are likely to go undetected [8].

From our discussions with hospital infection control practitioners, we discovered that low support low confidence rules may be interesting. The reason is simple: For a period of one year, patients' records come in thousands. Each attribute may have hundreds of (for example, the attending doctor) or tens of (for example, the organism code) values. Their combinations are like several thousand different products in the supermarket case. There can be billions or even trillions of possible association rules. An unexpected increase in a particular event can be easily lost among these various possible patterns, since the event's support will be very low. Thus, by setting a relatively high support threshold, potential interesting patterns are likely to go undetected. In our research, we set the support and confidence thresholds to 0.1%, so that we avoided the chance of missing interesting unusual patterns. Since our system is focused on finding interesting rules that may have been ignored by existing association rule induction systems, we further restrict our search to association rules with support less than 1% and confidence less than 30%. Thus, only low support and low confidence rules are analyzed.

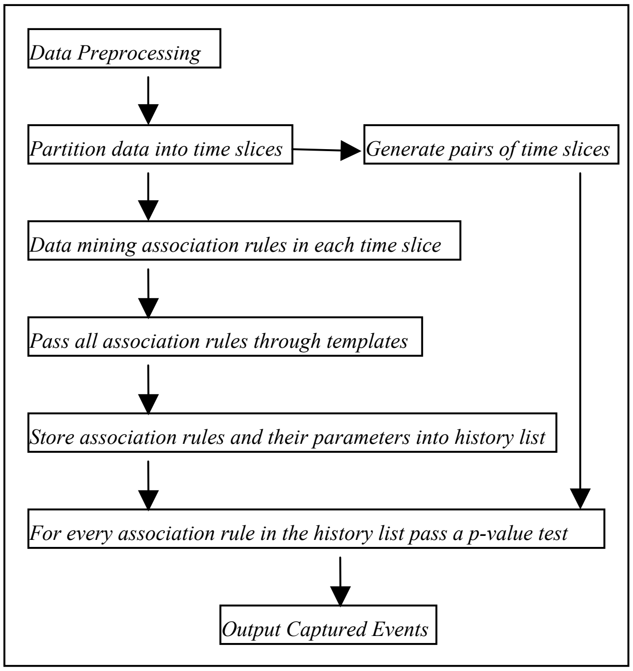
Data Preprocessing
Partition data into time slices
Generate pairs of time slices
Data mining association rules in each time slice
Pass all association rules through templates
Store association rules and their parameters into history list
For every association rule in the history list pass a p-value test
Output Captured Events

Figure 4: System Flow Chart

Figure 4 is the flow chart of our current system. After data preprocessing, we divide the data into different time slices, according to users' specification. We use apriori algorithm to discover all association rules that satisfy the predefined support and confidence thresholds and then pass these rules through various templates. Only association rules that passed all the templates (might be interesting to the users) will be stored in the history list, which is organized by time order. At the same time, we generate all possible time pairs out of these time slices. According to these pairs of time slices, every association rule in the history list will undergo a p-value test. Association rules which pass the test will be output as captured events.

RESULTS

Within the study period, a total of 49 low support (0.1% <support< 1%), low confidence (0.1% <confidence< 30%) events were captured. The expert reviewer was easily able to inspect all of them in less than a half-hour. From our discussion with infection control experts, they found 37 out of the 49 events (75.51%) encode interesting patterns, which may have potential significance for their research. Representative examples of the different classes of events are shown in figure 5.

drugcode=OXA, hospital=14, gender=M, organismspecies=STAPHYLOCOCCUS, -> resultcode=R

	Confidence	***Support***
June	*28.6%*	*0.1%*
July	*5.9%*	*0.1%*

Meaning: a significant decrease in clindamycin-resistant coagulase-positive Staphylococcus among male patients in hospital 14, from June 2000 to July 2000.

organismspecies=ENTEROCOCCUS, gender=M, drugcode=AM, agegroup=4 -> resultcode=R

	Confidence	***Support***
June	*13.3%*	*0.1%*
July	*28.6%*	*0.1%*

Meaning: a significant increase in Ampicillin-resistant Gram Positive Cocci Enterococcus among male patients age 60 and up, from June 2000 to July 2000.

drugcode=CFEP, hospital=1 organismspecies=PSEUDOMONAS -> resultcode=I

	Confidence	***Support***
May	*10%*	*0.2%*
June	*22.2%*	*0.2%*

Meaning: a significant increase in Cefepime intermediate resistant Pseudomonas in hospital 1, from May 2000 to June 2000.

Figure 5: Sample Output

After data preprocessing (eliminating duplicates, removing uninteresting attributes, and categorizing age and species into a small number of categories), 99% of uninteresting patterns were filtered out, and system run time was reduced from one hour to a more reasonable several seconds per time slice pair. Furthermore our three additional association rule templates in Figure 2 reduced the number of rules in the history by an additional 70%.

DISCUSSION

The results of this study indicate that low-support, low-confidence rules may have significant utility for infection control surveillance. This result means that in addition to mining for high-support, low-confidence association rules described in [4], it may also be important to mine for low-confidence, low-support association rules.

Based on the discussion with infection control practitioners, they found that although not all events were interesting, some suggested potential nosocomial outbreaks and changes of patterns in microbial resistance. This approach also makes it easy for experts to inspect events that might otherwise be missed by usual (manual) infection control surveillance methods.

Because a large number of low-support, low-confidence association rules can be found even in small data sets, a successful implementation of the process depends on efficient algorithms and on data selection and preprocessing strategies. We took several approaches to improving the efficiency of our association-rule mining system. First, we used the previously developed apriori algorithm. Second, we performed significant data pre-processing to reduce the search space by 99%. Lastly, we employed a set of rule templates to filter out rules of less importance or no importance to infection control practitioners. In doing so, we were able to mine a set of microbiology culture data spanning three months (using a one-month time slice) in a reasonable amount of time.

CONCLUSIONS

In this study, we employed new criteria (low-support, low-confidence) to automatically identify new, unexpected, and potentially interesting patterns in hospital infection control. The application of these new criteria raised a significant issue of efficiency of the approach to rule mining. To address this issue, we used the fast apriori algorithm, applied additional rule templates and performed data preprocessing. We found that low support, low-confidence rules are likely to have value for infection control surveillance. Furthermore, our approach to efficiency achieved reasonable running times.

FUTURE WORK

Future work includes formal evaluation of the rules generated by our approach. One simple approach is to randomly permute the data sets and check if the association rule method finds any interesting rules given the same templates. We also want to further explore whether the actions taken by infection control practitioners in response to such rules can reduce costs, morbidity, and mortality of nosocomial infections.

REFERENCE

[1] Kameniaca S, Cosic G, Lukic N, Miladinov-Mikov M: Hospital-acquired infections at the Institute of Oncology. Archive of oncology 2000; 8(3): 185-6.

[2] Centers for Disease Control and Prevention. Public health focus surveillance: prevention and control of nosocomial infections. Morbidity and mortality weekly report 1992; 41: 783-7.

[3] Samore M, Lichtenberg D, Saubermann L, Kawachi C, Carmeli Y: A clinical data repository enhances hospital infection control.

[4] Brossette SE, Sprague AP, Jones WT, Moser SA: A data mining system for infection control surveillance. Methods of information in medicine 2000; 39: 303-10.

[5] Agrawal R, Srikant R: Fast Algorithms for mining association rules.

[6] Borgelt C, Kruse R: Induction of association rules: apriori implementation.

[7] Panackal AA, M'ikanatha NM, Tsui F-C, McMahon J, Wagner MM, Dixon BW, Zubieta J, Phelan M, Mirza S, Morgan J, Jernigan D, Pasculle AW, Rankin JT, Hejjeh RA, Harrison LH. Automatic electronic laboratory-based reporting of notifiable infectious diseases at a large health system. Emerging Infectious Diseases 8(7):685-691, 2002.

[8] Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA: Association rules and data mining in hospital infection control and public health surveillance. Journal of the American medical informatics association 1998; 5: 373-381.

ACKNOLOGMENT

We thank Dr. Jeremy Espino and Ms. Patricia Harris for providing valuable information. This work was supported by grants GO8 LM06625-01, and T15 LM/DE07059 from the National Library of Medicine; Defense Advanced Research Projects Agency; contract 290-00-0009 from the Agency for Healthcare Research and Quality; Pennsylvania Department of Health Award number ME-01-737. This paper was also supported by Cooperative Agreement Number U90/CCU318753-01 and UP0/CCU318753-02 from the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.